Evaluating Auroral Forecasts Against Satellite Observations

M. K. Mooney[1,2], M. S. Marsh[2], C. Forsyth [1], M. Sharpe [2], T. Hughes [2], S. Bingham [2], D. R. Jackson [2], I. J. Rae [3], G. Chisham [4]

[1]Mullard Space Science Laboratory, University College London, Surrey UK
[2]Met Office, Exeter, UK
[3]Northumbria University, Newcastle, UK
[4]British Antarctic Survey, Cambridge, UK

Key Points:

- The OVATION-Prime 2013 nowcast model used at the UK Met Office is compared with auroral boundaries from IMAGE FUV using forecasting metrics.
- As a deterministic forecast, the OVATION-Prime 2013 nowcast predicts the location of the auroral oval well with a ROC score of 0.82.
- As a probabilistic forecast, the OVATION-Prime 2013 nowcast tends to underpredict the occurrence of the aurora by a factor of 1.1 - 6.

Corresponding author: Michaela K. Mooney, m.mooney.16@ucl.ac.uk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2020SW002688.

Abstract

The aurora is a readily visible phenomenon of interest to many members of the public. However, the aurora and associated phenomena can also significantly impact communications, ground-based infrastructure and high altitude radiation exposure. Forecasting the location of the auroral oval is therefore a key component of space weather forecast operations. A version of the OVATION-Prime 2013 auroral precipitation model [*Newell et al.*, 2014] was used by the UK Met Office Space Weather Operations Centre (MOSWOC). The operational implementation of the OVATION-Prime 2013 model at the UK Met Office delivered a 30-minute forecast of the location of the auroral oval and the probability of observing the aurora. Using weather forecast evaluation techniques, we evaluate the ability of the OVATION-Prime 2013 model forecasts to predict the location and probability of the aurora occurring by comparing the forecasts with auroral boundaries determined from data from the IMAGE satellite between 2000 and 2002. Our analysis shows that the operational model performs well at predicting the location of the auroral oval, with a relative operating characteristic (ROC) score of 0.82. The model performance is reduced in the dayside local time sectors (ROC score = 0.59) and during periods of higher geomagnetic activity (ROC score of 0.55 for Kp=8). As a probabilistic forecast, OVATION-Prime 2013 tends to under-predict the occurrence of aurora by a factor of 1.1 - 6, while probabilities of over 90% are over-predicted.

1 Plain Language Summary

Enhanced auroral activity at Earth can cause disruption to long-range radio communications and ground induced currents making forecasting the location of the auroral oval and probability of the aurora occurring of interest to many sectors such as aviation, energy and defence. The UK Met Office uses a version of the OVATION-Prime 2013 auroral forecast model to deliver a 30-minute forecast of the location and probability of observing the aurora. In this study, we evaluate the performance of the auroral forecasts against satellite observations of the aurora, captured by the IMAGE satellite between 2000-2002. Our analysis shows that the auroral forecast model performs well at predicting the location of the auroral oval, under nominal space weather conditions, but the probabilities of aurora occurring forecast by the model tend to be underpredicted, in other words, the aurora occurs more frequently than the forecast model predicts.

2 Introduction

Particles in the magnetosphere can be subject to acceleration or scattering processes resulting in particles being lost to the upper atmosphere. Particles precipitating into the upper atmosphere undergo collisions which result in a cascade of free electrons. These free electrons undergo further collisions, losing energy until they can eventually collisionally excite atmospheric atoms and ions. The resulting de-excitation emits a photon of radiation which we observe as aurora at altitudes of $\sim$ 100 km.

Forecasting the location and intensity of the aurora is of interest to many stakeholder industries such as the aviation, defense and energy sectors [*Cannon et al.*, 2013]. The free electrons and excited molecules in the upper atmosphere are known to degrade long-range radio communications in ultra-high frequency (UHF) wavebands [*Moore*, 1951; *Harang and Stroffregen*, 1940; *Jones et al.*, 2017]. Radio wave scattering can cause radar backscatter, resulting in radar clutter [*Elkins*, 1980; *Jones et al.*, 2017], and can also result in broadband noise in radio receivers [*Benson and Desch*, 1991; *Jones et al.*, 2017]. Increased electron precipitation in the upper atmosphere can also cause increased absorption of radio signals in the ionosphere [*Greenberg and LaBelle*, 2002; *Jones et al.*, 2017]. Ionospheric currents associated with enhanced auroral activity can induce currents in the ground which can damage ground-based infrastructure

such as electricity supply networks e.g. [*Erinmez et al.*, 2002; *Cannon et al.*, 2013; *Freeman et al.*, 2019; *Smith et al.*, 2019]. In addition, forecasting the occurrence of visible aurora is of importance for auroral tourism and is a key tool in promoting public awareness and engagement with space weather, through projects such as Aurorasaurus [*MacDonald et al.*, 2015] and AuroraWatch UK [*Case et al.*, 2017].

The auroral oval is highly dynamic with activity driven by factors both internal (e.g. geomagnetic substorms) and external to the magnetosphere (e.g., the interaction with the solar wind). Prolonged periods of southward directed interplanetary magnetic field can increase the open flux content of the magnetosphere which causes the auroral oval to expand to lower latitudes [*Cowley and Lockwood*, 1992]. During substorms, the sudden onset of reconnection in the magnetotail leads to a rapid brightening and widening in the nightside auroral oval which spreads eastwards and westwards during the substorm expansion phase [e.g. *Akasofu*, 1964].

The OVATION auroral forecast model [*Newell et al.*, 2002, 2010a] is an empirical model which predicts the location of the auroral oval based on the upstream solar wind conditions. The most recent version, OVATION-Prime 2013 (OP-2013) [*Newell et al.*, 2014] uses average particle precipitation maps obtained from Defense Meteorological Space Program (DMSP) satellites [*Hardy et al.*, 1984, 1985] spanning 21 years between 1 January 1984 to 31 December 2005, UV auroral data from the Global Ultraviolet Imager (GUVI) instrument onboard the TIMED satellite and real time solar wind conditions measured at the L1 point to produce maps of the predicted auroral flux. A version of the OP-2013 auroral forecast model has been implemented in daily operations of leading space weather forecasting centres including the U.S. National Oceanic Atmospheric Administration Space Weather Prediction Center (SWPC), the U.S. Department of Defense, Space Weather Operations Centre [*Jones et al.*, 2017] and the UK Met Office. The OP-2013 model was originally supplied to the Met Office by SWPC, however the operational implementations at the Met Office and SWPC have since diverged.

Forecast evaluation is an important step in both the implementation and development of space weather forecast models. Model verification can provide information on the skill, accuracy and reliability of models and also provides quantitative benchmarks to compare different forecast models. Previous verification studies have evaluated the performance of the earlier generation aurora forecast model, OVATION-Prime 2010 [*Newell et al.*, 2010a,b; *Machol et al.*, 2012; *Mitchell et al.*, 2013; *Lane et al.*, 2015; *Kosar et al.*, 2018]. *Newell et al.* [2010b] and *Machol et al.* [2012] evaluated the auroral forecasts of OP-2010 against ultraviolet images of the aurora from the instruments onboard the Polar satellite. *Newell et al.* [2010b] compared the instantaneous and hourly averaged predicted auroral power to the observed power estimated from Polar UVI data. The auroral power predicted by OP-2010 was found to be correlated with the observed auroral power from Polar UVI with a correlation coefficient $r^2 = 56\%$ for the instantaneous power forecast and an $r^2 = 58\%$ for the hourly averaged auroral power, demonstrating that just over half of the observed auroral power can be forecast by the OP-2010 model. *Mitchell et al.* [2013] found that OP-2010 described 47% of the variance in the Polar UVI nightside auroral power while a similar auroral model, OVATION-SM which uses averaged DMSP precipitation maps and ground magnetometer data from SuperMAG, described 71% of the nightside variance. *Machol et al.* [2012] used binary event analysis to evaluate the forecasts from OP-2010 and the suitability of the model as a tool for forecasting visible nightside aurora. *Machol et al.* [2012] compared the nightside auroral forecast to the boundaries derived from a fixed brightness threshold of the nightside auroral emission in the Polar UVI data. The result of this verification study found that the OP-2010 had a hit rate of 0.58 (the proportion of correct positive forecasts out of the total positive observations of aurora), a false alarm rate of 0.14 (the proportion of aurora forecasts which were not observed)

and an overall accuracy of 0.86 (the proportion of correct positive and negative forecasts over the total number of forecasts). *Lane et al.* [2015] performed a comparison study of the energy flux outputs forecast from 3 different models: OP-2010, the Kp-based auroral forecast model by *Hardy et al.* [1991], and a ring current model from the Space Weather Modeling Framework [*Fok et al.*, 2001; *Tóth et al.*, 2005]. Similarly to *Machol et al.* [2012], *Lane et al.* [2015] also used fixed thresholds to define the equatorward auroral boundary defined from particle precipitation measurements from the DMSP satellites. The authors presented the results in terms of the prediction efficiency, which is the model's ability to describe the percentage variance in the observed data set. The prediction efficiencies of OP-2010 were found to be 0.55 and 0.58 for the threshold values of 0.4 $erg\ cm^{-2}\ s^{-1}$ and 0.6 $erg\ cm^{-2}\ s^{-1}$, respectively.

Verification is important in monitoring model performance and also acts as a benchmark against which proposed improvements to the model can be tested. Verification techniques that are routinely used in terrestrial weather forecasting are now being applied to space weather forecast models. Binary event analysis is a method of comparing model forecasts with a ground-truth observational dataset and is widely used in many applications. The approach of using binary event analysis has been applied to evaluate nowcast and forecast models for example, in the verification study of OP-2010 by *Machol et al.* [2012] and verification studies of other space weather models including predicting magnetopause crossings [*Lopez et al.*, 2007; *Welling and Ridley*, 2010], radiation belt models [*Ganushkina et al.*, 2015, 2019; *Forsyth et al.*, 2020], temporal changes in the induced ground magnetic field (dB/dt) [*Pulkkinen et al.*, 2013] and solar flare forecasts [*Barnes et al.*, 2016; *Kubo et al.*, 2017; *Leka et al.*, 2019; *Murray et al.*, 2017; *Sharpe and Murray*, 2017].

In this study, we present an evaluation of auroral forecasts from the version of OP-2013 that was being used operationally at the Met Office, until December 2020. We compare the auroral forecasts from the model against auroral boundaries derived by *Longden et al.* [2010] from global FUV images of the auroral oval obtained by the IMAGE satellite. In particular, and in contrast to *Machol et al.* [2012]; *Newell et al.* [2010b]; *Lane et al.* [2015], we examine the output auroral probabilities from the operational auroral forecast, rather than the physical quantities (the predicted auroral power, energy or auroral flux) provided by the underlying OP-2013 model. We assess the model performance in predicting the location of the auroral oval using binary event analysis and present the results in Relative Operating Characteristic (ROC) curves. We also assess the forecast probabilities of aurora occurring output by the model using reliability curves. Our results show that, overall, the model performs well at predicting the location of the auroral oval, but the forecast probabilities tend to under-predict auroral occurrence. Furthermore, we show that the model results are substantially less reliable on the dayside and during periods of enhanced geomagnetic activity.

3 Data and Evaluation Methods

3.1 Forecast Model: OP-2013

Both the OP-2010 and OP-2013 versions of the auroral forecast model [*Newell et al.*, 2009, 2010a, 2014] predict the precipitating electron and proton auroal flux based on upstream solar wind conditions, measured at L1. *Newell et al.* [2009] created averaged particle precipitation maps of the auroral oval collected by the SSJ instruments onboard the Defense Meteorological Space Program satellites (DMSP) and categorised the DMSP particle precipitation energy spectra into four categories of aurora: monoenergetic, broadband and diffuse electron aurora and ion aurora. *Newell et al.* [2009] determined a linear scaling between the electron and proton flux from the DMSP data with an empirically derived solar wind coupling function [*Newell et al.*, 2007]. The upstream solar wind data required includes the B_z and B_y components of the inter-

planetary magnetic field, the total magnetic field strength, the solar wind velocity and the IMF clock angle. In each model grid point, the particle flux was calculated as a function of season and the type of aurora. For OP-2013, additional UV auroral data from the Global Ultraviolet Imager (GUVI) instrument onboard the TIMED satellite is included to improve the performance of the model at higher values of Kp, between Kp 5 - 8 [*Newell et al.*, 2014]. The resultant maps of linear scaling coefficients are then used to predict the precipitating electron and proton fluxes under all upstream conditions. Additional improvements to the model made in the upgrade from OP-2010 to OP-2013 include further noise reduction and a smoother data interpolation in the post-midnight MLT sectors [*Newell et al.*, 2014]. We direct the interested reader to *Newell et al.* [2007, 2009, 2010a, 2014] for full details of the OP-2010 and 2013 models.

The Met Office operational implementation of the version of OP-2013 assessed in this study, assumes a fixed 30 minute propagation time for the solar wind measured at the L1 point to arrive at Earth. In this operational version, the combined precipitating particle flux from all types of aurora at each grid point is linearly scaled into an estimated probability of aurora occurring which is interpreted as the probability of an observer seeing the visible aurora. The linear conversion of auroral flux to probability implemented in the version of OP-2013 at the Met Office are as originally developed by SWPC and could be further refined. The forecast probabilities were tuned by SWPC in response to citizen science observations under the assumption that the forecast probabilities of aurora occurring were mainly used by members of the public and may under-predict the probability of aurora occurring (Rodney Viereck, private communications). The results of this study could be used to tune the forecast probability to optimise the forecast reliability. Further details on the conversion from the predicted auroral flux to the probability of aurora occurring is included in the Supplementary Information. The operational implementation provides an auroral forecast for both the northern and southern hemispheres 30 minutes ahead of the current time.

The original OP-2013 IDL code was supplied to the Met Office by SWPC. In 2016, the Met Office converted the code to Python and returned the Python version of OP-2013 to SWPC. In October 2020, SWPC implemented an upgraded version of OVATION termed *OVATION 2020* which, again, differs from the Met Office implementation. OVATION 2020 uses an improved geomagnetic field model to provide a more accurate auroral location. In addition, OVATION 2020 provides the modelled energy flux in $ergs/cm^2$ as well as the scaled probability of seeing the aurora. SWPC have also implemented an estimate of the solar wind driving based on Kp data to use as an alternative to run the model when upstream solar wind data is unavailable. Details of the SWPC auroral forecast using OVATION 2020 can be found on the SWPC website.

The version of the OP-2013 model evaluated in this study was used operationally at the Met Office until December 2020. The Met Office currently use an alternative Kp-driven 3-day forecast version of the OP-2013 model. We note that the Kp-driven version was developed at the Met Office independently of the SWPC Kp-driven model. The Met Office may return the 30-minute forecast version of OP-2013 evaluated in this study to operational use in the future to operate in parallel with the Kp-driven 3-day forecast version. In this paper, we refer to the 30 minute auroral forecast as a *nowcast* to distinguish it from the alternative 3-day auroral forecast which is currently in operation at the Met Office.

In this study, we produce hindcasts of the output from the 30-minute nowcast version of OP-2013 used at the Met Office using historic solar wind data for the period between May 2000 to October 2002, not auroral forecasts that were issued in near real time by the Met Office. Figure 1a shows an example output of the OP-2013 northern hemisphere 30 minute auroral forecast from 25 September 2000. The model output was produced using Advanced Composition Explorer (ACE) solar wind data. The

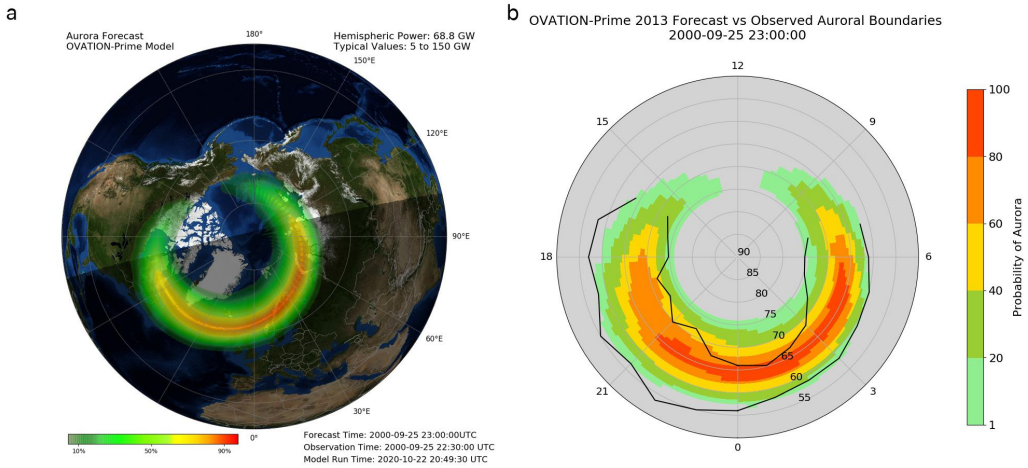

Figure 1. Panel (a) shows an example output forecast from OP-2013 showing the northern hemisphere auroral forecast 30 minutes ahead for 23:00 on the 25 September 2000, in geographic coordinates. The model output was produced using ACE solar wind data. The colour scale shows the probability of aurora occurring with green showing lower probability and red showing higher probabilities. The day/night terminator is indicated on the map as the line separating the dark and light faces of Earth and the estimated total hemispheric auroral power is shown in the top right hand corner. Panel (b) shows the OP-2013 forecast (colour shading) for the same date and time as in panel (a) but plotted in magnetic coordinates (magnetic latitude by magnetic local time (MLT)). The black lines show the equatorial and poleward boundaries of the aurora from *Longden et al.* [2010] for the forecast date and time. We note that in panel (a), the central meridian is centred on 2300 local time and in panel (b) the central meridian is centred on 0000 local time and so the contours are effectively rotated by a 1-hour MLT sector.

auroral oval is plotted on geographic coordinates with the colour scale showing the forecast probability of aurora occurring.

3.2 Observational Data: Auroral Boundaries Derived from IMAGE FUV Data

The NASA IMAGE satellite was in operation between 2000-2005 in a highly elliptical, precessing polar orbit which enabled it to capture images of the northern and southern polar regions. The orbit had an initial perigee of 1000 km and an apogee of 44000 km ($\sim 7R_e$) [*Mende et al.*, 2000a]. Between 2000-2002, the orbital apogee was situated over the northern hemisphere. IMAGE carried a far-ultraviolet (FUV) wideband imaging camera (WIC) sensitive to emission between 140-190 nm [*Mende et al.*, 2000b] which took images of the Earth approximately every 2 minutes, determined by the spin of the spacecraft [*Burch*, 2000].

Using IMAGE FUV data, *Longden et al.* [2010] developed an automated technique to identify the poleward and equatorward luminosity boundaries of the auroral oval. The IMAGE FUV data was converted from geomagnetic coordinates to altitude-adjusted corrected geomagnetic coordinates (AACGM) [*Baker and Wing*, 1989]. *Longden et al.* [2010] created a latitudinal intensity profile of auroral emission in each of the 24 magnetic local time (MLT) sectors and fitted these profiles with both single and double Gaussian profiles. The profile fits to the data were evaluated using the

reduced χ^2 statistic and the best fit function retained. The poleward and equatorward auroral luminosity boundaries (PALBs and EALBs, respectively) are defined as the poleward and equatorward the points on the Gaussian curve where the auroral intensity drops to half the peak value (the full width half maximums (FWHMs)) of the best fitting Gaussian function, offset from the centre of the Gaussian peak. We direct the interested reader to the full description of the method published in *Longden et al.* [2010]. The boundaries determined by *Longden et al.* [2010] provide a single location of the poleward and equatorward boundaries of the auroral oval in each MLT sector, without any assumption of the global shape of the auroral oval. Auroral boundaries were identified for each global auroral image, with a cadence of 2 minutes.

In this study, we use the poleward and equatorward auroral luminosity boundaries determined from the IMAGE WIC data by *Longden et al.* [2010] as a *ground truth* observational data set to compare with the model forecast probability maps output from OP-2013. The poleward boundary identifications from *Longden et al.* [2010] have been shown to be co-located with the poleward emission boundary measured from DMSP within 3°on average in all MLT sectors, making the boundaries a suitable observational dataset to compare with the OP-2013 forecasts. The auroral boundary data available for the northern auroral oval spanned 30 months from May 2000 to October 2002 [*Chisham*, 2017]. Figure 1b shows a comparison of the probability forecast maps from OP-2013 to the poleward and equatorward auroral boundaries determined by *Longden et al.* [2010] in MLT and magnetic latitude (MLAT) coordinates. The colours show the 30 minute forecast of the probability of aurora occurring as output from OP-2013. Grey regions indicate a forecast probability of aurora occurring of less than 1%. The black lines show the corresponding observed boundaries. We note that, in this example, there is a lack of observed auroral boundaries in some dayside MLT sectors. While the method of *Longden et al.* [2010] aims to identify the poleward and equatorward auroral luminosity boundaries in each MLT sector, the number of successful boundary identifications in dayside sectors is much lower than on the nightside [*Mooney et al.*, 2020]. The dayside aurora tends to be dimmer and thinner [*Holzworth and Meng*, 1975; *Carbary*, 2005] and is more contaminated with dayglow making it more difficult to identify the dayside auroral boundaries. In this study, we only evaluate the model where there are corresponding observational auroral boundaries.

3.3 Verification Method

In this study, we have produced the OP-2013 auroral forecasts spanning the period of May 2000 - October 2002 [*Marsh and Mooney*, 2021], coinciding with the available observational auroral boundary data from *Longden et al.* [2010], using historic solar wind data measured by the ACE satellite, provided by the National Oceanic Atmospheric Administration (NOAA). Each forecast requires four hours of input solar wind data, thus in order to ensure that the forecasts were independent of one another, we down-sampled our forecast dataset to four hour resolution. To match the model forecast and the observational ground truth auroral boundaries, we use the auroral boundaries that were closest in time and within ±2.5 minutes of the 4 hour separated forecast time. This resulted in 3360 corresponding forecast and observation pairs. The magnetic latitude (MLAT) range of the OP-2013 data spans 50 - 89.5 °and covers 24 hours of MLT, with a grid resolution of 0.25 MLT by 0.5 °MLAT.

In this evaluation study, we use two verification techniques that are widely used in terrestrial weather forecast verification. Firstly, we apply binary event analysis to evaluate how well the OP-2013 model discriminates between auroral and non-auroral regions via comparison with the *Longden et al.* [2010] boundaries. This evaluates how well the model performs as a deterministic forecast for predicting the location of the auroral oval. We test over a range of forecast probability levels, between 0 - 100% in 10% increments. At a particular level, for each available forecast and observation

pair in each grid cell with a forecast probability that exceeds the level, we determine whether or not the aurora was observed at that grid cell. We repeat this test to build up truth tables for different forecast probability thresholds. If the forecast probability of aurora occurring is equal to or greater than the set level and aurora was also observed, it counts as a *hit* in our truth table. If the forecast probability of aurora occurring is equal to or greater than the set level but aurora was not observed, it counts as a *false alarm*. If the forecast probability of aurora occurring is less than the set level and aurora was observed, it counts as a *miss*. If the forecast probability of aurora occurring is less than the set level but aurora was not observed, it counts as a *correct negative*. From the truth tables for each level, we evaluate the hit rate (hits/(hits+misses)) and false alarm rate (false alarms/(false alarms + correct negative forecasts)). These hit rates and false alarm rates are combined and presented on ROC curves [*Swets et al.*, 1955; *Swets*, 1988; *Mason*, 1982]. ROC curves are obtained by plotting the calculated hit rate against the false alarm rate from the truth table, for each 10% probability level. A ROC score, calculated as the fractional area under the ROC curve, provides a quantitative summary of the model discrimination indicated by the ROC plot. A ROC score between 0.5 - 1 indicates that the hit rate exceeds the false alarm rate for most probability levels and that the model is skillful in discriminating events from non-events.

Secondly, we assess the validity of the forecast probabilities against the observed occurrence of the aurora using reliability (or attribute) diagrams [*Jolliffe and Stephenson*, 2012; *Hsu and Murphy*, 1986; *Wilks*, 2006]. The forecast model would be completely reliable if, over all the occasions during the assessment period when the forecast probability was p, the aurora was observed p% of the time. However, if the forecast probabilities and observed frequencies of occurrence do not have a one-to-one correspondence, the reliability diagram provides information on whether the model is under-forecasting or over-forecasting the probabilities. This information can be used to re-calibrate the forecast probabilities by rescaling the probability of aurora occurring against the observed occurrence of aurora. We provide suggestions of how the forecast probabilities of aurora occurring may be adjusted based on the results of this study in Section 5.2. Attribute diagrams are similar to reliability diagrams, showing the observed frequency of an event against the forecast probabilities but they include additional information such as the average, climatology value of the observations and forecasts which can be used to assess the forecast model in more detail. Further detail on ROC and reliability analysis is provided in the Supplementary Information.

ROC and reliability analysis are standard methods used in forecast verification by the weather community (for example, *Dube et al.* [2017]). They have been used to evaluate flare forecasts from the Met Office Space Weather Operations Centre (MOSWOC) in studies by *Murray et al.* [2017] and *Sharpe and Murray* [2017], to evaluate the performance of a new radiation belt forecast model [*Forsyth et al.*, 2020] and to assess a sudden storm commencement probabilistic forecast model [*Smith et al.*, 2020].

The spherical geometry of the auroral forecasts means that the area of each grid cell is not uniform. This can influence how well the forecast is judged to perform. For example, near the pole, where aurora are not generally expected to occur, there is a greater concentration of grid cells than at 60°, where there is a greater likelihood of auroral activity. To account for this, the inputs into our ROC and reliability analysis were weighted by the cosine of the latitude of each grid cell e.g. [*Young*, 2010].

4 Results

In the following section, we present the results of our evaluation of the OP-2013 model using the locations of the auroral boundaries derived from IMAGE WIC data. In Section 4.1 we present the results of the analysis of 2.5 years of data between May 2000

and October 2002 in all MLT sectors. In Sections 4.2 and 4.3 we present the results of the verification during the four seasons of 2001 and in different MLT sectors around the auroral oval to test for seasonal and spatial variations in the forecast performance. In Section 4.4, we present the results of the verification during geomagnetically active times for different values of Kp.

4.1 Model Evaluation Between May 2000 - October 2002

Figure 2 shows the ROC curve from the comparison of $\sim$ 2.5 years of model forecast and observation pairs. The curve is constructed by setting the probability threshold in 10% increments to calculate the hit rates and false alarm rates. The ROC curve shown in Figure 2 shows that, over the 2.5 year verification period, the model performs well and has a ROC score (fractional area under the curve) of 0.82. At each 10% probability increment the model hit rate is higher than the false alarm rate with a maximum difference between the hit rate and false alarm rate of 0.6 for probabilities exceeding 5%. Thus the forecasts perform well at predicting the location of the aurora overall. The probability bin centred on 10% has the largest difference between the hit rate and the false alarm rate, also referred to as the Peirce Score [*Peirce*, 1884]. This shows that a probability of between 5 - 15% is the threshold at which the OP-2013 model performs the best at discriminating between regions of aurora and no aurora, compared to the observed auroral boundaries.

Figure 3 shows the reliability diagram for the full $\sim$ 2.5 year verification period, plotting the occurrence rate from auroral observations for given forecast probability ranges. Figure 3 shows that the aurora are largely under-predicted, with occurrence frequencies greater than the forecast probabilities for probabilities up to 80%. The lowest non-zero probabilities of 10% and 20% are under-predicted by a factor of $\sim$ 6 while the 80% probabilities are only under-predicted by a few percent. The 90% and 100% probability bins slightly over-predicted the probability of aurora occurring with the highest probability value of 100% over-predicting the occurrence by $\sim$ 20%, a factor of 1.25.

The dotted horizontal and vertical lines indicate the observed climatological frequency of occurrence of aurora is 0.30, calculated as the fraction of positive auroral observations that the aurora did occur out of the total number of auroral observations. The histogram in Figure 3 shows the number of data points in each forecast probability bin. The histogram shows that the probabilities forecast by the OP-2013 model are distributed across all probability bins and are not clustered around the climatology value. The lowest forecast probability bin contains all forecasts issued with a probability of 5% and lower and has the largest number of data points. This bin is dominated by the grid points where the main auroral oval is rarely or never predicted to occur, for example at low and high magnetic latitudes. The large number of forecasts with a low probability of aurora occurring in this bin correspond to a large number of observations where the aurora was not observed to occur which reduces the overall observed climatology (mean occurrence). The solid pink diagonal line of no skill lies mid-way between the diagonal line of perfect reliability and the horizontal climatology line. Points on the reliability curve which lie above/below the line of no skill, contribute positively/negatively to the Brier skill score. Pink shading indicates the region where the forecast is skilful compared with the in-sample climatology. The majority of the points on the reliability line lie in the shaded skill region except for probabilities of 10% and 20% which appear to be extremely under-predicted by the OP-2013 model. The Brier skill score [*Brier*, 1950; *Murphy*, 1973] of -0.03 indicates that overall, the OP-2013 model is not more skilful at predicting when the aurora occurs than simply always forecasting the within-sample climatology of 0.30. While the Brier skill score indicates that the OP-2013 model is not more skillful than using a climatological forecast, the attributes diagram shows that the majority of forecast probabilities are

skillful. The discrepancy in the conclusions drawn from these two analyses metrics highlights the increased understanding of the model performance that can be gained from using the full attributes diagram rather than only using value of the Brier skill score.

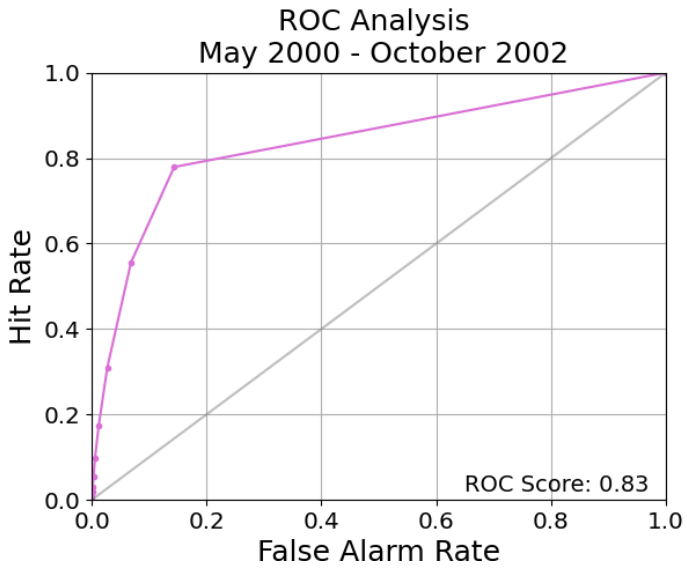

Figure 2. The result of the ROC analysis from the 2.5 years of model and observation comparisons. Each point on the ROC curve corresponds to the hit rate vs false alarm rate in each 10% threshold bin. The high ROC score of 0.82, defined by the fractional area under the ROC curve, shows that the model performs well at predicting the location of the auroral oval.

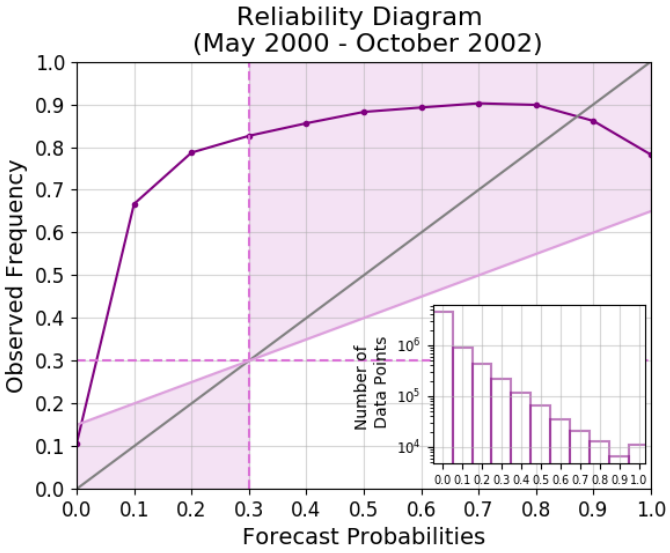

Figure 3. Reliability diagram showing the the results from the comparison of 2.5 years of auroral forecasts against observations. The histogram shows the distribution of the forecast probabilities over the 2.5 year period. The grey diagonal line indicates the perfect reliability line of 1:1 correspondence between the forecast probabilities and the observed aurora. Regions where the pink reliability line lies above/below the grey diagonal line indicate that the model is under-forecasting/over-forecasting the occurrence of aurora. The vertical and horizontal dashed lines show the observed climatology. The solid pink diagonal line of no skill delineating the shaded region, lies mid-way between the diagonal line of perfect reliability and the horizontal climatology line. Data points within the shaded region contribute positively to the Brier skill score.

4.2 Seasonal Verification During 2001

Seasonal variations in ionospheric conductivity as a result of the solar zenith angle affects the auroral precipitation [*Liou et al.*, 2001; *Newell et al.*, 1996, 2010a]. The seasonal variation in the auroral emission was examined by *Newell et al.* [2010a] and implemented in the OP-2013 model by calculating the the predicted auroral flux as a function of season. Here, we have evaluated the seasonal variability in the model performance. For the seasonal analysis we use data and forecasts from 5 February 2001 into 4 February 2002 as this is the only complete year of WIC observational data including all seasons. The seasons were defined similarly to the those used by *Newell et al.* [2010a] as being 90 days centred on the equinoxes and solstices. The start and end dates of each season were then adjusted slightly to include the 6 uncategorised days that fall between the seasons by this definition. The seasonal dates used in the analysis are as follows: spring is between 5 of February to the 6 of May; summer is between 7 May to 8 August; autumn is between 9 August and 6 November; winter is between 7 November and 4 February.

Figure 4a shows the ROC curves for each season in 2001 - 2002. There is some seasonal variation in the ROC scores in Figure 4a, with ROC scores ranging from 0.79 - 0.86 however these scores are similar and indicate that the model performs well in identifying the auroral oval in all seasons. The results of the ROC scores for each full season between May 2000 - October 2002 are provided in the Supplementary Information. On average, the spring and winter ROC scores are consistently highest,

with spring and winter seasons having mean ROC scores of 0.86, while the summer ROC scores were the lowest, with a mean of 0.77 across the three summer periods.

The seasonal variation in the ROC score may be indicative of the model performance but it may also be due to the seasonal variations in the identification of the auroral boundaries. During the summer months, the increased UV contamination from reflected sunlight reduces the number of successfully identified auroral boundaries in the WIC data. We also note that the ROC scores of summer and autumn 2002 are reduced compared to the same season in previous years. In summer 2002, the IMAGE satellite suffered damage to the boom which affected the satellite pointing and resulting in an increased uncertainty in spacecraft pointing [*Frey*, 2010] and thus increased uncertainty and thus in the location of the auroral boundaries.

Figure 4 shows the reliability diagram for each of the seasons in 2001. The seasonal reliability is consistent with the overall reliability shown in Figure 3. In all seasons, the occurrence frequency increases rapidly with probability, thus there is an under-prediction of the auroral occurrence. For autumn and spring forecasts, the observed auroral occurrence plateaus at ~ 0.8 and ~ 0.9 respectively for mid range forecast probabilities between $\sim 20 - 70\%$, whereas the occurrence rate in summer and winter increases steadily with probability above a forecast probability of 20%. The autumn forecasts are over-predicted at the higher probability values of 70% and above. We note that there is no significant difference in the solar or geomagnetic activity between the seasons in 2001. Generally, there are a higher number of auroral boundary observations to compare with the OP-2013 model forecasts in winter, as indicated by the winter histogram in Figure 4, however this is not expected to have a considerable effect on the verification results.

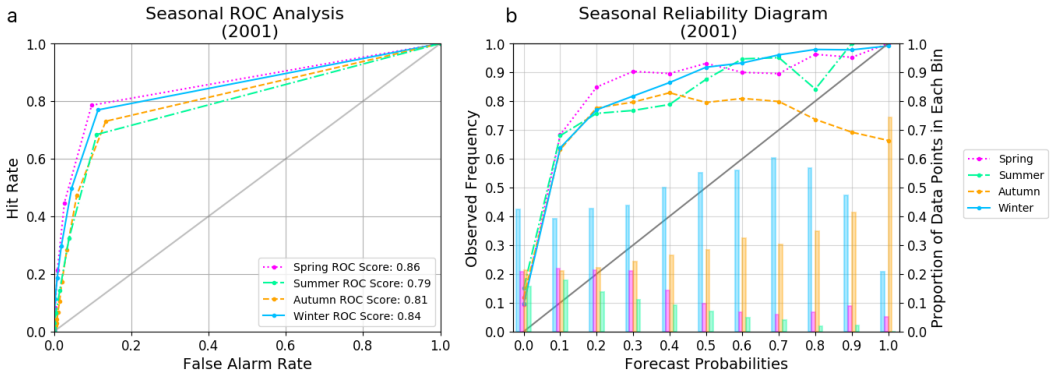

Figure 4. (a) The results of the ROC analysis for each season in 2001. The high ROC scores for each season demonstrate that the OP-2013 model performs well all year round. (b) The results of the reliability analysis for each season in 2001. The histogram shows the proportion of data in each season, for each probability bin. In both panels, the results for spring, summer, autumn and winter are shown by dotted pink, dot-dash green, dashed orange and solid blue lines, respectively.

4.3 MLT Dependence

The shape of the auroral oval varies with MLT sector. Typically, the dayside auroral oval tends to be thinner and dimmer [*Holzworth and Meng*, 1975; *Carbary*, 2005] while the nightside aurora generally extends over a wider magnetic latitude range and is more variable with enhanced auroral precipitation linked to magnetospheric

activity such as substorms. Here, we evaluate the performance of the OP-2013 model in the noon, dawn, dusk and midnight regions. Each region is defined as three hours of MLT centred on MLT sectors 00, 12, 06 and 18. The ROC curves of each 3-hour MLT sector are shown in Figure 5 and show that the model performs well in the dawn, dusk and midnight sectors, with ROC scores of between 0.78-0.86. However, the ROC score from the noon region is considerably lower, at 0.59 showing that forecast model does not perform as well in this region. Using a probability threshold of 10% to indicate the presence of aurora only gives a hit rate of ∼0.2, much lower than the hit rates of 0.6-0.85 seen in the other MLT sectors. The results in the truth table for the noon analysis are dominated by missed forecasts and correct rejections where the aurora is not forecast by the model. The lack of forecast aurora in this region may be because of a data gap in the underlying DMSP particle precipitation data, due to the dawn-dusk orbit of the spacecraft. The midnight data gap was interpolated over in the upgrades between OP-2010 and OP-2013 [*Newell et al.*, 2010a], however there are no details on whether the corresponding dayside data gap was interpolated.

Figure 5b shows the reliability diagrams for each three hour MLT region. The reliability curves for the dawn, dusk and midnight sectors are similar to those of the 2.5 year verification shown in Figure 3 with forecast probabilities below 70 - 80% being largely under-predicted and greater than 80% being over-predicted. The reliability diagram from the noon MLT sectors is quite different to the other MLT sectors. The reliability curve from the noon MLT sectors shows that the OP-2013 model tends to under-predict when forecasting aurora with probabilities less than 30% and over-predict when forecasting aurora with probabilities between 30% and 60%; whereas, aurora was not forecast with probabilities >70%.

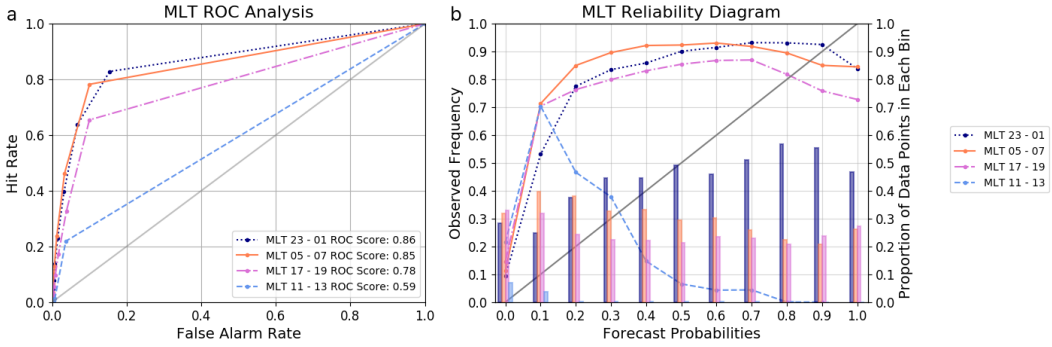

Figure 5. (a) A ROC analysis for 4 longitudinal regions of the auroral oval. (b) A reliability analysis for 4 longitudinal regions of the auroral oval, each spanning 3 hours of MLT. The MLT sectors for midnight (23-01), dawn (05-07), dusk (17-19) and noon (11-13) and are shown by dotted navy, solid orange, dot-dashed pink and dashed light blue respectively.

4.4 Kp Dependence

In the following section, we evaluate the performance of the OP-2013 aurora forecast model under different levels of geomagnetic activity based on Kp. Kp levels of 5 and above are generally considered to be geomagnetically active periods and so it is important to evaluate the performance of the OP-2013 model during these levels of geomagnetic activity which can have a real impact on daily services at Earth. The OP-2010 model was known to break down at higher levels of geomagnetic activity of Kp $\geq$ 5 [*Newell et al.*, 2014]. This led to the inclusion of additional GUVI data at higher Kp levels (Kp 5 - 8) as part of the upgrade to the OP-2013 generation.

All corresponding forecast and observation pairs between May 2000 and October 2002 were divided into subsets based on the level of Kp measured at the time. The results of the ROC analysis, including all the ROC scores for each Kp level, are shown in Figure 6. The ROC scores generally decrease for increasing levels of Kp, with Kp = 1 having a ROC score of 0.83 and Kp = 8 having a ROC score of 0.55. The ROC scores for Kp = 1 - 6 are within 0.05 of each other, implying that the model performs relatively well at discriminating between auroral and non-auroral regions at these levels of activity. However, at the highest activity levels of Kp = 7 and Kp = 8, the ROC score drops to 0.7 and 0.55 respectively. While these ROC scores indicate that the forecast has some skill in identifying where the aurora will be, these forecasts are less skillful than at lower activity levels. The results for Kp = 8 show that the hit rates are lower and the false alarm rates are higher compared to the results for lower Kp levels, indicating that the model is predicting that aurora will occur but not always in the correct locations, compared to the observed auroral boundaries. It is not uncommon for models which perform well within a nominal range of average conditions to not perform as well during extreme events. Despite the improvements made to the OP-2013 model using GUVI data during higher levels of Kp, these observations are likely limited due to the rarity of periods of extremely high Kp. It would be informative to repeat this analysis with the predecessor OP-2010 model to quantify the improvement made by including the GUVI data at high levels of Kp.

Figure 6b and c shows the reliability diagrams for Kp levels of 1-8. The reliability curves for Kp levels 1-5 plateau at an observed frequency of ∼0.8 - 0.9 for forecast probabilities of 30% and above. The reliability curves for Kp levels 6-7 plateau at a lower observed frequency of aurora of ∼ 0.7 - 0.8 for forecast probabilities of 10% and above. Kp = 8 shows the reliability curve dropping with increasing probability such that the observed occurrence of high probabilities is much lower than the forecast probability indicating a more concerning over-prediction. From the histogram, we note that Kp levels between 1-3 are the most common, with the highest number of points in these categories representing low geomagnetic activity. Kp levels of 7 and 8 are statistically much more rare events and have the lowest number of data points in the ROC and reliability analysis. The inclusion of more data in the analysis for this level of high geomagnetic activity would help to confirm this evaluation of the OP-2013 model at these high Kp levels.

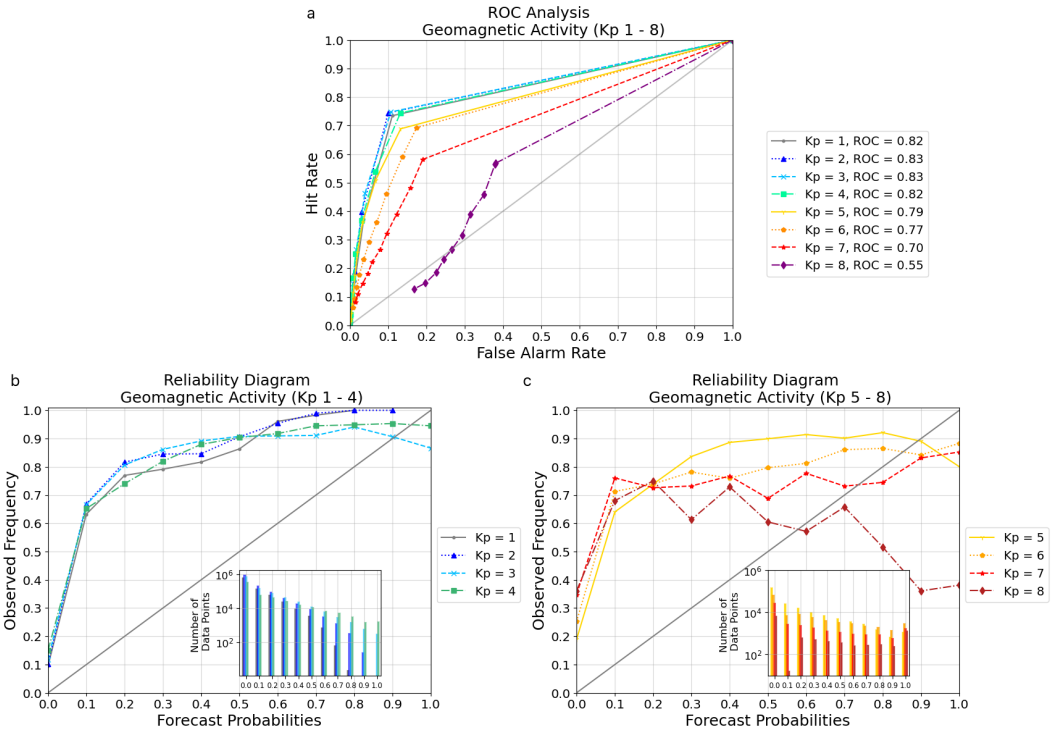

Figure 6. (a) The results of the ROC analysis for the OP-2013 model during different levels of geomagnetic activity spanning Kp = 1 - 8. (b) The reliability diagram for the OP-2013 model during different levels of geomagnetic activity spanning Kp = 1 - 4. (c) The reliability diagram for the OP-2013 model during different levels of geomagnetic activity spanning Kp = 5 - 8.

5 Discussion

In this study, we have used auroral boundaries derived from global IMAGE FUV data between May 2000 to October 2002 to evaluate the performance of the auroral forecasts made by the OP-2013 model, used operationally at the Met Office. Using a combination of ROC and reliability analysis, we find that overall, the OP-2013 model performs well at predicting the location of the aurora with a ROC scores of between 0.70 to 0.86, although the forecast skill was notably lower around noon (ROC score of 0.59) and at higher Kp (ROC score of 0.55, for Kp = 8). The overall ROC score compares well with other space weather forecasts, such as M-class solar flare forecasts [*Murray et al.*, 2017]. The OP-2013 forecast probabilities tended to under-predict the occurrence of the aurora, with the observation frequency of the aurora typically plateauing at ∼0.8 for forecast probabilities exceeding 20%.

5.1 Deterministic Auroral Forecasts

The results of the ROC analysis show that overall, the model performs well as a deterministic model at discriminating between regions of aurora and no aurora. In the seasonal analysis, while there is some seasonal variability in ROC scores, all ROC scores are greater than 0.74, indicating that the model performs well year-round.

In the evaluation of OP-2013 by MLT sector, the model had a lower ROC score in the dayside MLT sectors centred on the noon MLT (11-13 MLT). The noon MLT sectors had a ROC score of 0.59 compared to the dusk (17-19 MLT), dawn (05-07

MLT) and midnight (23-01 MLT) sectors which had ROC scores between 0.78 - 0.86. The higher ROC scores in the nightside MLT sectors indicate that the model performs better at predicting the location of the aurora in these regions.

The width of the auroral oval varies with local time sector. The nightside auroral oval is typically wider and more dynamic than the dayside. The nightside auroral dynamics are primarily driven by substorms which cause a rapid expansion and brightening of the auroral oval. The solar wind driven OP-2013 model is unable to forecast substorm activity and the 30 minute resolution of the operational forecasts cannot capture substorm dynamics, however the change in width of the auroral oval during substorms occurs within the average predicted auroral oval location. *Mooney et al.* [2020] showed that during substorms, the poleward boundary of the auroral oval moves by up to 3°in the substorm onset MLT sectors. During substorms, the typical width of the auroral oval varies by 10 - 17°[*Walach et al.*, 2017]. Compared to the width of the auroral oval, the 3°change in the poleward boundary represents a small change of 17 - 30% of the total oval width. In addition, after the substorm activity has subsided, the auroral oval generally returns to the same size and width that it had prior to the substorm and so substorms have no lasting effect on the auroral oval. During a substorm, the relatively small expansion in the oval width in the substorm onset sectors near midnight would result in a slight increase in the number of missed forecasts but this does not have a big impact on the overall ROC score. While the OP-2013 model cannot forecast if or when a substorm may occur, the occurrence of a substorm has a relatively low impact on the performance of the OP-2013 model. In contrast, the lower ROC scores of the model in the noon sectors indicate that the model does not forecast the location of the dayside auroral oval particularly well. In the dayside MLT sectors, the auroral oval is generally much thinner and so any offset between the observed and forecast locations of the auroral oval will result in a bigger reduction in the overall ROC score.

In the final part of this study in Section 4.4, we focused on the performance of OP-2013 during periods of different levels of geomagnetic activity, defined by Kp level. Overall, the ROC scores decrease with increasing Kp levels from 0.83 to 0.55 for Kp = 1 and Kp = 8, respectively. All ROC scores for Kp 1 - 7 are greater than 0.70. The relatively high ROC scores above 0.70 for geomagnetic activity up to Kp = 7 may indicate that the additional GUVI data is having a positive effect on the performance of OP-2013 at disturbance levels between Kp 5 - 7. It would be informative to repeat this analysis to evaluate the performance of OP-2010 during high Kp levels to confirm and quantify the improvement made by including GUVI data. The low ROC score of 0.55 for Kp = 8 is likely due to the rarity of periods of extremely high Kp and thus the model is less well constrained. This could suggest that the linear scaling of auroral flux with solar wind driving used by *Newell et al.* [2007] to construct the OP-2010 and OP-2013 models breaks down during more extreme and statistically more rare events of Kp $\geq$ 7.

5.2 Evaluating the Forecast Auroral Probabilities

The reliability diagrams show that the forecast probabilities of aurora occurring tend to be under-predicted, that is that the aurora occurs more frequently than the model predicts, particularly for lower probability values of less than 80%. At the highest forecast probability values, greater than 80%, the model tends towards a slight over-prediction of the probability of aurora occurring. This is observed in most cases from the seasonal, MLT sector and geomagnetic activity analysis.

The observed frequency of aurora does not increase linearly with the forecast probabilities but instead is relatively constant between 0.8 - 0.9 for for all the forecast-observation pairs for forecast probabilities of 20% and above. This means that the lower

forecast probabilities of 20% are under-predicted by a factor of ∼6. As the forecast probability of aurora occurring tends towards the observed frequency of aurora, the difference between the forecast probability and the observed frequency decreases and so the factor of how much the aurora is under or over-predicted also decreases.

The results of the reliability analysis show that the conversion from auroral flux to probability of aurora occurring is not particularly robust, however, this conversion is a non-trivial task. Using the results of the reliability analysis, a correction to re-calibrate the probabilities forecast by the model could be developed to improve the reliability OP-2013 auroral forecasts. The probabilities forecast by the OP-2013 model vary with season, MLT sector and geomagnetic activity (Kp level) which would need to be accounted for if a correction were to be developed. However, the results of the reliability analysis showed that for forecast probabilities of above 20 - 30%, the observed occurrence of aurora is approximately constant at around 0.7-0.9, which would make it difficult to linearly re-scale the forecast probabilities. All forecast probabilities of $\sim$ 20% or above would be re-scaled to an $\sim$ 80% probability of aurora occurring, effectively producing a deterministic forecast.

Germany et al. [1998] found that the brightness of the UV electron aurora is proportional to the total electron energy flux with a conversion factor of approximately 0.12 R per $erg\ cm^{-2}\ s^{-1}$. This conversion was utilised by *Machol et al.* [2012] and *Case et al.* [2017] to define the poleward and equatorward boundaries from the Polar UVI images and OP-2010 output. The conversion of the total electron energy flux to brightness could be used to derive a more robust method of converting the predicted auroral fluxes into a probability of aurora occurring. However, given the difficulties of linearly scaling the auroral flux into a probability of aurora occurring, it may be preferable to develop a flux threshold system, using the conversion of *Germany et al.* [1998]. For example, in regions where the predicted auroral flux is greater than zero indicates that there may be some auroral effects. In regions where the auroral flux exceeds a certain brightness threshold would indicate that the aurora should be visible and the brightest aurora would be predicted in the regions of maximum auroral flux.

5.3 Comparisons with Previous Auroral Forecast Evaluation Studies

Newell et al. [2010b]; *Machol et al.* [2012]; *Lane et al.* [2015] evaluated the auroral forecasts from OP-2010. From these three studies, the binary event analysis methods applied by *Machol et al.* [2012] are the most comparable to the analysis applied in this study. *Machol et al.* [2012] evaluated the use of the OP-2010 model as an operational forecast model for visible aurora by assessing the deterministic ability of the model to forecast the location of the aurora compared to UVI observations. We have similarly examined how well the OP-2013 model performs as a deterministic forecast of the location of the aurora, although using IMAGE FUV data as our ground truth and utilising the ROC curves and scores to examine the performance of the model. Extending this, we have also examined the validity of the forecast probabilities of aurora occurring as well as examining the performance of the model with season, local time and geomagnetic activity.

The most notable difference between the analysis presented in this study and the analysis of *Machol et al.* [2012], other than the updated model in this study, is the determination of the ground truth data. *Machol et al.* [2012] compared the locations of model predictions of electron fluxes exceeding 1 $erg\ cm^{-2}\ s^{-1}$ and auroral luminosities from Polar UVI exceeding 0.25 kR whereas we used auroral luminosity boundaries determined from IMAGE WIC data by *Longden et al.* [2010]. As such, a direct comparison between the results cannot be used to infer any change in performance between the OP-2010 and OP-2013 models, but may still be informative.

Table 1 shows the verification statistics calculated from the 10%, 50% and 80% binary event analysis in this study with the results of *Machol et al.* [2012]. In the study by *Machol et al.* [2012], the results were presented in terms of the false alarm ratio (as defined by *Wilks* [2006]). In Table 1 we present our results for the false alarm rate and the false alarm ratio, for comparison with the results of *Machol et al.* [2012]. The equations for all the verification statistics in Table 1 are provided in the Supporting Information. Comparing the results of *Machol et al.* [2012] and the 10% bin from this analysis, all of the statistics are within 15%. *Machol et al.* [2012] found that by increasing the energy flux threshold used to define the location of the auroral boundaries, resulted in an increase in the number of false positives and a decrease in the number of false negatives in the truth table. The binary event analysis presented here, similarly shows that as the probability threshold is increased, the number of false positives increases and the number of false negatives decreases.

Overall, the results of the verification statistics from both studies show a similar performance for both the OP-2010 and OP-2013 generations of the model. We caution that the results of these two studies cannot be directly compared to assess improvements made between the two generations of the model. Differences in the results between the two studies presented in Table 5.3 may reflect the upgrades made to the model between the OP-2010 and OP-2013 generations, however due to the differences in the observational datasets and the definition of the observed auroral boundaries between this study and the study by *Machol et al.* [2012], the comparison of the two sets of results cannot be used to quantify the upgrades implemented in the model.

Table 1. A comparison of the verification statistics derived from the results of the 10%, 50% and 80% probability thresholds from current OP-2013 evaluation study presented in this paper with those from the OP-2010 evaluation study carried out by *Machol et al.* [2012].

Verification Statistic	**Results from *Machol et al.* [2012] Analysis of OP-2010**	**Results from Present Analysis of OP**		
		10%	**50%**	**80%**
Hit Rate	58%	73%	8%	2%
False Alarm Rate	–	11%	0%	0%
False Alarm Ratio	14%	25%	11%	14%
Proportion of True Positives	86%	75%	89%	86%
Proportion of False Negatives	26%	12%	30%	31%
Accuracy	77%	84%	71%	69%

6 Conclusions

In this study we have evaluated the performance of the version of OP-2013 that was used operationally by the Met Office in daily space weather forecasts by comparing the forecast outputs with the location of the auroral oval identified from IMAGE FUV data by *Longden et al.* [2010]. We have applied forecast evaluation techniques which are routinely used in terrestrial weather forecast verification to assess both the deterministic and probabilistic nature of the auroral forecast model. Overall, the OP-2013 model performed well at predicting the location of the auroral oval, with ROC scores of between 0.70 to 0.86, although the forecast skill was notably lower around noon (ROC score of 0.59) and at higher Kp (ROC score of 0.55, for Kp = 8). The reliability analysis showed that the observed frequency of aurora is constant at 80 - 90% for forecast probabilities of $\sim$ 20% and above and does not scale linearly with increasing forecast probability. This results in the lower forecast probabilities of 20% being significantly under-predicted, by a factor of 6 i.e. the aurora occurs 6 times more frequently than the model predicts for a forecast probability of 20%. The highest forecast probabilities of $\sim$ 90-100% are over-predicted by up to approximately 20%; that is the aurora occurs up to 20% less frequently than the model predicts for these high forecast probability values. The majority of forecast probabilities are skilful with the exception of the 10% and 20% probabilities which are substantially under-predicted. The results of the reliability analysis from this study could be used to recalibrate the forecast probabilities of aurora occurring and improve the Met Office auroral forecasts.

The ROC and reliability analysis presented in this study show a robust methodology that is widely used in terrestrial weather forecast verification that can also be applied to a wide range of space weather forecast models which have an appropriate set of observations to use in the analysis. These methods can be used to fairly compare forecasts from similar models or to quantify improvements made to space weather models during model development. The results presented in this analysis provide a performance benchmark against which upgrades to the OP-2013 auroral forecast model or alternative auroral forecast models can be fairly and quantitatively tested. Our analysis also highlights the further insight into the reliability of the forecast probabilities of aurora occurring output by the model from using attribute diagrams in addition to calculating the Brier skill score, compared to solely using the Brier skill score.

Acknowledgments

Michaela Mooney is supported by a UK NERC NPIF studentship 1926376. Colin Forsyth is supported by NERC Independent Research Fellowship NE/N014480/1. Jonathan Rae and Colin Forsyth are supported by the STFC Consolidated Grant to MSSL ST/S000240/1. Gareth Chisham is supported as part of the BAS Polar Science for Planet Earth Programme, funded by the UK Natural Environment Research Council (NERC) as part of United Kingdom Research and Innovation (UKRI). This work was carried out in partnership with the Met Office. The auroral hindcast dataset produced from the Met Office operational implementation of the Ovation-Prime 2013 nowcast model that was used in this study was provided by the Met Office and is available at: http://doi.org/10.5281/zenodo.4653288. The Ovation Prime 2013 code was provided to the Met Office by Rodney Viereck of the NOAA Space Weather Prediction Center and the original Ovation Prime code was developed by Patrick Newell and colleagues at Johns Hopkins University. Historic solar wind data from the ACE satellite was provided by Douglas Biesecker at the National Oceanic Atmospheric Administration and are available at https://sohoftp.nascom.nasa.gov/sdb/goes/ace/monthly/. Auroral boundary data were derived and provided by the British Antarctic Survey based on IMAGE satellite data (https://www.bas.ac.uk/project/image-auroral-boundary-data/). These boundary data are freely available from doi.org/10.5285/75aa66c1-

47b4-4344-ab5d-52ff2913a61e. The IMAGE FUV data were provided courtesy of the instrument PI Stephen Mende (University of California, Berkeley). We thank the PI, the IMAGE mission, and the IMAGE FUV team for data usage and processing tools. IMAGE FUV data is archived at: https://cdaweb.gsfc.nasa.gov/pub/data/image/fuv/. Kp index data was provided by ISGI GFZ Potsdam. Archive Kp index data is available at: https://spaceweather.gfz-potsdam.de/products-data/nowcasts/nowcast-kp-index/downloads. We thank Rodney Viereck at SWPC and Edmund Henley at the Met Office for useful discussion.

References

Akasofu, S.-I. (1964), The development of the auroral substorm, *Planetary and Space Science*, *12*, 273–282, doi:10.1016/0032-0633(64)90151-5.

Baker, K. B., and S. Wing (1989), A new magnetic coordinate system for conjugate studies at high latitudes, *Journal of Geophysical Research*, *94*(A7), 9139–9143, doi:10.1029/JA094iA07p09139.

Barnes, G., K. D. Leka, C. J. Schrijver, T. Colak, R. Qahwaji, O. W. Ashamari, Y. Yuan, J. Zhang, R. T. J. McAteer, D. S. Bloomfield, P. A. Higgins, P. T. Gallagher, D. A. Falconer, M. K. Georgoulis, M. S. Wheatland, C. Balch, T. Dunn, and E. L. Wagner (2016), A Comparison of Flare Forecasting Methods. I. Results from the "All-Clear" Workshop, *Astrophysical Journal*, *829*(2), 89, doi:10.3847/0004-637X/829/2/89.

Benson, R. F., and M. D. Desch (1991), Wideband noise observed at ground level in the auroral region, *Radio Science*, *26*(4), 943–948, doi:10.1029/91RS00450.

Brier, G. W. (1950), Verification of Forecasts Expressed in Terms of Probability, *Monthly Weather Review*, *78*, doi:10.1175/1520-0493(1950)078¡0001:VOFEIT¿2.0.CO;2.

Burch, J. L. (2000), IMAGE mission overview, *Space Science Reviews*, *91*, 1–14.

Cannon, P., M. Angling, L. Barclay, C. Curry, C. Dyer, R. Edwards, G. Greene, M. Hapgood, R. Horne, D. Jackson, C. Mitchell, J. Owen, A. Richards, C. Rogers, K. Ryden, S. Saunders, M. Sweeting, R. Tanner, A. Thomson, and C. Underwood (2013), Extreme space weather: Impacts on engineered systems and infrastructure, *Tech. rep.*, Royal Academy of Engineering.

Carbary, J. F. (2005), A Kp-based model of auroral boundaries, *Space Weather*, *3*(10), S10001, doi:10.1029/2005SW000162.

Case, N. A., S. R. Marple, F. Honary, J. A. Wild, D. D. Billett, and A. Grocott (2017), AuroraWatch UK: An Automated Aurora Alert System, *Earth and Space Science*, *4*(12), 746–754, doi:10.1002/2017EA000328.

Chisham, G. (2017), Auroral boundary derived from image satellite mission data (may 2000 - oct 2002), version 1.1, Polar Data Centre, Natural Environment Research Council, UK., doi:10.5285/75aa66c1-47b4-4344-ab5d-52ff2913a61e.

Cowley, S. W. H., and M. Lockwood (1992), Excitation and decay of solar wind-driven flows in the magnetosphere-ionosphere system, *Annales Geophysicae*, *10*(1-2), 103–115.

Dube, A., R. Ashrit, H. Singh, K. Arora, G. Iyengar, and E. N. Rajagopal (2017), Evalutating the performance of two global ensemble forecasting sytems in predicting rainfall over india during the southwest monsoons, *Meteorological Applications*, *24*, 230–238, doi:10.1002/met.1621.

Elkins, T. J. (1980), A model for high frequency radar auroral clutter, *Tech. rep.*, Rome Air Development Institution.

Erinmez, I. A., J. G. Kappenman, and W. A. Radasky (2002), Management of the geomagnetically induced current risks on the national grid company's electric power transmission system, *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*(5-6), 743–756, doi:10.1016/S1364-6826(02)00036-6.

Fok, M. C., R. A. Wolf, R. W. Spiro, and T. E. Moore (2001), Comprehensive computational model of Earth's ring current, *Journal of Geophysical Research*, *106*(A5), 8417–8424, doi:10.1029/2000JA000235.

Forsyth, C., C. E. J. Watt, M. K. Mooney, I. J. Rae, S. D. Walton, M. Marsh, R. B. Horne, and J. Albert (2020), Forecasting GOES 15 ¿ 2MeV electron fluxes from solar wind data and geomagnetic indices, *Space Weather*, doi:10.1029/2019SW002416.

Freeman, M. P., C. Forsyth, and I. J. Rae (2019), The Influence of Substorms on Extreme Rates of Change of the Surface Horizontal Magnetic Field in the United Kingdom, *Space Weather*, *17*(6), 827–844, doi:10.1029/2018SW002148.

Frey, H. (2010), Image fuv log.

Ganushkina, N. Y., O. A. Amariutei, D. Welling, and D. Heynderickx (2015), Nowcast model for low-energy electrons in the inner magnetosphere, *Space Weather*, *13*(1), 16–34, doi:10.1002/2014SW001098.

Ganushkina, N. Y., I. Sillanpää, D. Welling, J. Haiducek, M. Liemohn, S. Dubyagin, and J. V. Rodriguez (2019), Validation of Inner Magnetosphere Particle Transport and Acceleration Model (IMPTAM) With Long-Term GOES MAGED Measurements of keV Electron Fluxes at Geostationary Orbit, *Space Weather*, *17*(5), 687–708, doi:10.1029/2018SW002028.

Germany, G. A., J. F. Spann, G. K. Parks, M. J. Brittnacher, R. Elsen, L. Chen, D. Lummerzheim, and M. H. Rees (1998), Auroral observations from the POLAR ultraviolet imager (UVI), in *Geospace Mass and Energy Flow: Results From the International SolarTerrestrial Physics Program*, vol. 104, edited by L. Horwitz, L. Gallagher, and K. Peterson, pp. 149–160, Geophys. Monogr. Ser., doi:10.1029/GM104p0149.

Greenberg, E. M., and J. LaBelle (2002), Measurement and modeling of auroral absorption of HF radio waves using a single receiver, *Radio Science*, *37*(2), 1022, doi:10.1029/2000RS002550.

Harang, L., and W. Stroffregen (1940), Echoversuche auf Ultrakurzwellen, *Hochfreq. Elektroakust*, *55*, 105–108.

Hardy, D. A., L. K. Schmitt, M. S. Gussenhoven, F. J. Marshall, and H. C. Yeh (1984), Precipitating electron and ion detectors (SSJ/4) for the block 5D/Flights 6-10 DMSP (Defense Meteorological Satellite Program) satellites: Calibration and data presentation, Unknow.

Hardy, D. A., M. S. Gussenhoven, and E. Holeman (1985), A statistical model of auroral electron precipitation, *Journal of Geophysics Research*, *90*, 42294248, doi:10.1029/JA090iA05p04229.

Hardy, D. A., W. McNeil, M. S. Gussenhoven, and D. Brautigam (1991), A statistical model of auroral ion precipitation. 2. of the average patterns, *Journal of Geophysical Research*, *96*(A4), 5539–5547, doi:10.1029/90JA02451.

Holzworth, R. H., and C. I. Meng (1975), Mathematical representation of the auroral oval, *Geophysical Research Letters*, *2*(9), 377–380, doi:10.1029/GL002i009p00377.

Hsu, W., and A. H. Murphy (1986), The attributes diagram A geometrical framework for assessing the quality of probability forecasts, *International Journal of Forecasting*, *2*(3), 285–293, doi:10.1016/0169-2070(86)90048-8.

Jolliffe, I. T., and D. B. Stephenson (2012), *Forecast verification: A practitioner's guide in atmospheric science*, 2 ed., Wiley.

Jones, J., S. Sanders, B. Davis, C. Hedrick, E. J. Mitchell, and J. M. Cox (2017), Research to Operations Transition of an Auroral Specification and Forecast Model, in *Advanced Maui Optical and Space Surveillance (AMOS) Technologies Conference*, p. 94.

Kosar, B. C., E. A. MacDonald, N. A. Case, Y. Zhang, E. J. Mitchell, and R. Viereck (2018), A case study comparing citizen science aurora data with global auroral boundaries derived from satellite imagery and empirical models, *Journal of Atmospheric and Solar-Terrestrial Physics*, *177*, 274–282, doi:

10.1016/j.jastp.2018.05.006.

Kubo, Y., M. Den, and M. Ishii (2017), Verification of operational solar flare forecast: Case of Regional Warning Center Japan, *Journal of Space Weather and Space Climate*, *7*, A20, doi:10.1051/swsc/2017018.

Lane, C., A. Acebal, and Y. Zheng (2015), Assessing predictive ability of three auroral precipitation models using DMSP energy flux, *Space Weather*, *13*(1), 61–71, doi:10.1002/2014SW001085.

Leka, K. D., S.-H. Park, K. Kusano, J. Andries, G. Barnes, S. Bingham, D. S. Bloomfield, A. E. McCloskey, V. Delouille, D. Falconer, P. T. Gallagher, M. K. Georgoulis, Y. Kubo, K. Lee, S. Lee, V. Lobzin, J. Mun, S. A. Murray, T. A. M. Hamad Nageem, R. Qahwaji, M. Sharpe, R. A. Steenburgh, G. Steward, and M. Terkildsen (2019), A Comparison of Flare Forecasting Methods. II. Benchmarks, Metrics, and Performance Results for Operational Solar Flare Forecasting Systems, *Astrophysical Journal, Supplement*, *243*(2), 36, doi:10.3847/1538-4365/ab2e12.

Liou, K., P. T. Newell, and C. I. Meng (2001), Seasonal effects on auroral particle acceleration and precipitation, *Journal of Geophysics Research*, *106*(A4), 5531–5542, doi:10.1029/1999JA000391.

Longden, N., G. Chisham, M. P. Freeman, G. A. Abel, and T. Sotirelis (2010), Estimating the location of the open-closed magnetic field line boundary from auroral images, *Annales Geophysicae*, *28*(9), 1659–1678, doi:10.5194/angeo-28-1659-2010.

Lopez, R. E., S. Hernandez, M. Wiltberger, C. L. Huang, E. L. Kepko, H. Spence, C. C. Goodrich, and J. G. Lyon (2007), Predicting magnetopause crossings at geosynchronous orbit during the Halloween storms, *Space Weather*, *5*(1), 01005, doi:10.1029/2006SW000222.

MacDonald, E. A., N. A. Case, J. H. Clayton, M. K. Hall, M. Heavner, N. Lalone, K. G. Patel, and A. Tapia (2015), Aurorasaurus: A citizen science platform for viewing and reporting the aurora, *Space Weather*, *13*(9), 548–559, doi:10.1002/2015SW001214.

Machol, J. L., J. C. Green, R. J. Redmon, R. A. Viereck, and P. T. Newell (2012), Evaluation of OVATION Prime as a forecast model for visible aurorae, *Space Weather*, *10*, doi:10.1029/2011SW000746.

Marsh, M. S., and M. K. Mooney (2021), OVATION-PRIME-2013 Met Office Nowcast Verification Dataset, doi:10.5281/zenodo.4653288.

Mason, I. (1982), A Model for Assessment of Weather Forecasts, *Australian Meteorological Magazine*, *30*, 291–303.

Mende, S. B., H. Heetderks, H. U. Frey, M. Lampton, S. P. Geller, S. Habraken, E. Renotte, C. Jamar, P. Rochus, J. Spann, S. A. Fuselier, J. C. Gerard, R. Gladstone, S. Murphree, and L. Cogger (2000a), Far ultraviolet imaging from the IMAGE spacecraft. 1. System design, *Space Science Reviews*, *91*, 243–270.

Mende, S. B., H. Heetderks, H. U. Frey, M. Lampton, S. P. Geller, R. Abiad, O. H. W. Siegmund, A. S. Tremsin, J. Spann, H. Dougani, S. A. Fuselier, A. L. Magoncelli, M. B. Bumala, S. Murphree, and T. Trondsen (2000b), Far ultraviolet imaging from the IMAGE spacecraft. 2. Wideband FUV imaging, *Space Science Reviews*, *91*, 271–285.

Mitchell, E. J., P. T. Newell, J. W. Gjerloev, and K. Liou (2013), OVATION-SM: A model of auroral precipitation based on SuperMAG generalized auroral electrojet and substorm onset times, *Journal of Geophysical Research (Space Physics)*, *118*(6), 3747–3759, doi:10.1002/jgra.50343.

Mooney, M. K., C. Forsyth, I. J. Rae, G. Chisham, M. S. Marsh, D. R. Jackson, S. Bingham, and B. Hubert (2020), Examining Local Time Variations the Gains and Losses of Open Magnetic Flux During Substorms, *Journal of Geophysics Research: Space Physics*, doi:10.1029/2019JA027369.

Moore, R. K. (1951), A VHF Propagation Phenomenon Associated with Aurora, *Journal of Geophysics Research*.

Murphy, A. H. (1973), A New Vector Partition of the Probability Score., *Journal of Applied Meteorology*, *12*(4), 595–600, doi:10.1175/1520-0450(1973)012¡0595:ANVPOT¿2.0.CO;2.

Murray, S. A., S. Bingham, M. Sharpe, and D. R. Jackson (2017), Flare forecasting at the Met Office Space Weather Operations Centre, *Space Weather*, *15*(4), 577–588, doi:10.1002/2016SW001579.

Newell, P. T., C.-I. Meng, and K. M. Lyons (1996), Suppression of discrete aurorae by sunlight, *Nature*, *381*(6585), 766–767, doi:10.1038/381766a0.

Newell, P. T., T. Sotirelis, J. M. Ruohoniemi, J. F. Carbary, K. Liou, J. P. Skura, C. I. Meng, C. Deehr, D. Wilkinson, and F. J. Rich (2002), OVATION: Oval variation, assessment, tracking, intensity, and online nowcasting, *Annales Geophysicae*, *20*(7), 1039–1047, doi:10.5194/angeo-20-1039-2002.

Newell, P. T., T. Sotirelis, K. Liou, C. I. Meng, and F. J. Rich (2007), A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables, *Journal of Geophysics Research (Space Physics)*, *112*(A1), A01206, doi:10.1029/2006JA012015.

Newell, P. T., T. Sotirelis, and S. Wing (2009), Diffuse, monoenergetic, and broadband aurora: The global precipitation budget, *Journal of Geophysics Research (Space Physics)*, *114*(A9), A09207, doi:10.1029/2009JA014326.

Newell, P. T., T. Sotirelis, and S. Wing (2010a), Seasonal variations in diffuse, monoenergetic, and broadband aurora, *Journal of Geophysics Research (Space Physics)*, *115*(A3), A03216, doi:10.1029/2009JA014805.

Newell, P. T., T. Sotirelis, K. Liou, A. R. Lee, S. Wing, J. Green, and R. Redmon (2010b), Predictive ability of four auroral precipitation models as evaluated using Polar UVI global images, *Space Weather*, *8*(12), S12004, doi:10.1029/2010SW000604.

Newell, P. T., K. Liou, Y. Zhang, T. Sotirelis, L. J. Paxton, and E. J. Mitchell (2014), OVATION Prime-2013: Extension of auroral precipitation model to higher disturbance levels, *Space Weather*, *12*(6), 368–379, doi:10.1002/2014SW001056.

Peirce, C. S. (1884), The numerical measure of the success of predictions, *Science*, *4*, 453 – 454.

Pulkkinen, A., L. RastäTter, M. Kuznetsova, H. Singer, C. Balch, D. Weimer, G. Toth, A. Ridley, T. Gombosi, M. Wiltberger, J. Raeder, and R. Weigel (2013), Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations, *Space Weather*, *11*(6), 369–385, doi:10.1002/swe.20056.

Sharpe, M. A., and S. A. Murray (2017), Verification of Space Weather Forecasts Issued by the Met Office Space Weather Operations Centre, *Space Weather*, *15*(10), 1383–1395, doi:10.1002/2017SW001683.

Smith, A. W., M. P. Freeman, I. J. Rae, and C. Forsyth (2019), The Influence of Sudden Commencements on the Rate of Change of the Surface Horizontal Magnetic Field in the United Kingdom, *Space Weather*, *17*(11), 1605–1617, doi:10.1029/2019SW002281.

Smith, A. W., I. J. Rae, C. Forsyth, D. M. Oliveira, M. P. Freeman, and D. R. Jackson (2020), Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning, *Space Weather*, doi:10.1029/2020sw002603.

Swets, J. A. (1988), Measuring the Accuracy of Diagnostic Systems, *Science*, *240*(4857), 1285–1293, doi:10.1126/science.3287615.

Swets, J. A., W. P. Tanner, and B. T. G. (1955), The evidence for a decision-making theory of visual detection, *Tech. rep.*, Engineering Research Institute, University of Michigan.

Tóth, G., I. V. Sokolov, T. I. Gombosi, D. R. Chesney, C. R. Clauer, D. L. de Zeeuw, K. C. Hansen, K. J. Kane, W. B. Manchester, R. C. Oehmke, K. G. Powell, A. J. Ridley, I. I. Roussev, Q. F. Stout, O. Volberg, R. A. Wolf, S. Sazykin, A. Chan, B. Yu, and J. Kóta (2005), Space Weather Modeling Framework: A new tool

for the space science community, *Journal of Geophysical Research (Space Physics)*, *110*(A12), doi:10.1029/2005JA011126.

Walach, M.-T., S. E. Milan, K. R. Murphy, J. A. Carter, B. A. Hubert, and A. Grocott (2017), Comparative study of large-scale auroral signatures of substorms, steady magnetospheric convection events, and sawtooth events, *Journal of Geophysical Research (Space Physics)*, *122*(6), 6357–6373, doi:10.1002/2017JA023991.

Welling, D. T., and A. J. Ridley (2010), Validation of SWMF magnetic field and plasma, *Space Weather*, *8*(3), 03002, doi:10.1029/2009SW000494.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2 ed., Elsevier.

Young, R. M. B. (2010), Decomposition of the brier score for weighted forecast-verification pairs, *Quarterly Journal of the Royal Meteorological Society*, *136*(650), doi:10.1002/qj.641.